פרויקט בקורס מבוא לניתוח נתונים – עמרי סביר וגל תמר

שאלה 5

בחנו במאגר הנתונים על מחלות לב.

שאלות המחקר

בשאלת בדיקת ההשערות (שאלה 3) בחרנו לחקור האם יש הבדל בגיל הממוצע בקרב המטופלים שיש להם מחלות לב בין גברים לנשים. בחרנו בשאלה זו מכיוון שעבור שנינו, על בסיס ידע קודם שהיה לנו בתחום, התשובה המיידית לשאלה זו הייתה שאכן יש הבדל והגיל הממוצע של הגברים נמוך יותר מהגיל הממוצע של הנשים, ורצינו לבחון האם הנתונים הללו מתיישבים עם מה שחשבנו.

בשאלת הסיווג (שאלה 4) רצינו ליצור ולאמן מסווג שיוכל על בסיס שאר הנתונים לקבוע האם למטופל מסוים יש מחלות לב או לא. בחרנו בשאלה זו כיוון שרצינו לדעת האם יש דרך לחזות ברמת דיוק גבוהה למי יש מחלות לב, בהתבסס על הנתונים האחרים שנבדקו.

מסד הנתונים

מסד הנתונים שקיבלנו הכיל 1025 רשומות על מטופלים שנבדקו בשנת 1988 ב- 4 בתי חולים שונים בקליבלנד, שוויץ, הונגריה ולוס אנג'לס. הערה: בהמשך הורדנו כפילויות ונשארנו עם 302 רשומות.

הקטגוריות שנאספו על כל מטופל הן:

age – גיל, משתנה נומרי
sex – מין, משתנה בוליאני
cp– סוג כאב החזה שהמטופל חווה, משתנה קטגוריאלי
trestbps – לחץ דם במנוחה, משתנה נומרי
chol – רמת כולסטרול, משתנה נומרי
fbs -רמת סוכר בדם אחרי צום, משתנה בוליאני
restecg – קצב לב, משתנה קטגוריאלי
thalach – דופק מקסימלי, משתנה נומרי
exang – כאב ראש אחרי פעילות גופנית, משתנה בוליאני
oldpeak – מקטע ST, משתנה נומרי
slope – תוצאת אקו-לב, משתנה קטגוריאלי
ca – מספר כלי הדם העיקריים שקיימת בהם בעיה בתפקוד, משתנה נומרי
thal – מחלות תורשתיות גנטיות במערכת הדם, משתנה קטגוריאלי
target – האם למטופל יש מחלות לב, משתנה בוליאני

הערה: מסד הנתונים שקיבלנו לא הכיל חוסרים בנתונים. עם זאת, במסד הנתונים הופיעו רשומות כפולות, כלומר רשומות זהות לחלוטין שחזרו על עצמן יותר מפעם אחת.

מכיוון שראינו שישנן רשומות אשר חוזרות על עצמן מספר פעמים, וההסתברות שב-14 פרמטרים שונים, הנתונים של "מטופלים" שונים יהיו זהים היא קטנה מאוד, תהינו האם עלינו להוריד שכפולים אלו. לכן, חיפשנו באינטרנט על מאגר נתונים זה, ומצאנו מאגר נתונים דומה (כנראה המקור) המכיל את אותן הרשומות שמופיעות במאגר הנתונים שלנו, אך בלי כל השכפולים. לכן, בסופו של דבר, מכיוון שלדעתנו ההסתברות שבמאגר נתונים זה השכפולים הינם טעות אנוש ולא נתונים אמיתיים היא גבוהה מאוד, החלטנו להוריד את כל השכפולים הללו.

קישור למקור: https://archive.ics.uci.edu/ml/datasets/heart+Disease

במידה שלא היינו מורידים את השכפולים, כאשר הם באמת שכפולים ולא נתונים זהים של מטופלים אחרים, אלגוריתם הסיווג היה נותן לנו את התוצאה K=1 מכיוון שלרוב הרשומות קיים לפחות שכפול אחד עודף, ולכן השכפול יעיד לבדו הכי טוב על המטופל אותו מנסים לסווג, לכן- תוצאות המסווג במקרה זה יהיו מוטות.

מסד הנתונים מתאים לחקר השאלות שבחרנו כיוון שהוא מכיל את כל הנתונים שאנחנו חייבים על מנת לחקור את הנושא בשאלת בדיקת ההשערות. בנוסף מסד הנתונים מכיל נתונים רבים על המטופלים; הן נתונים אישיים כמו גיל, מין ורקע גנטי והן נתונים רפואיים שנבדקו על המטופל כמו רמת כולסטרול בדם, רמת סוכר בדם ודופק. לכן אנו סבורים כי המגוון הרחב של הנתונים יאפשר לנו להגיע למסקנות מדויקות בשאלת החיזוי.

<u>הערה:</u> כלל הגרפים והניתוחים מתייחסים לדאטה פריים שלא כולל את מה שהחשבנו כשכפולים.

<u>ניתוח הנתונים וממצאים</u>

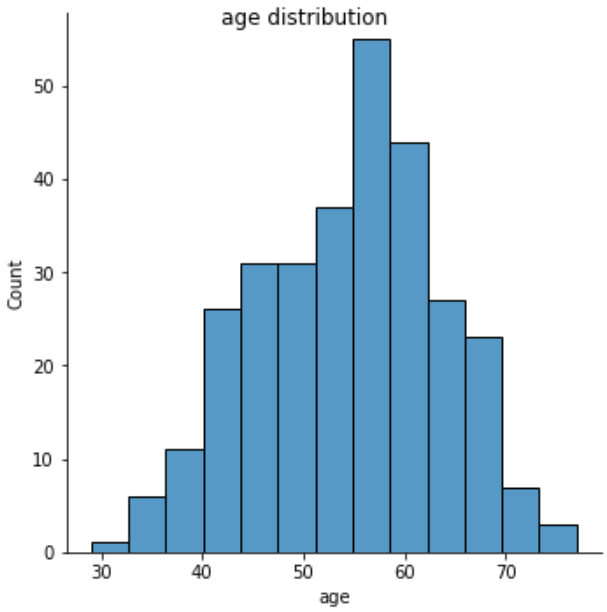
Exploratory data analysis

(שאלה 2 סעיף א')

בשאלה זו התבקשנו להציג את ההתפלגות של 5 משתנים. המשתנים שבחרנו להציג הם: age, sex, cp, target, thalach. בחרנו במשתנים אלו מכיוון שהם משתנים מרכזיים שאיתם נעבוד בשאלות הבאות. המשתנים age, sex, target רלוונטיים לשאלה 3 של בדיקת ההשערות. בשאלה 4 המשתנה target הוא המשתנה אותו נרצה לחזות ו- cp, thalach הם משתנים בהם נעזר בתהליך הסיווג.

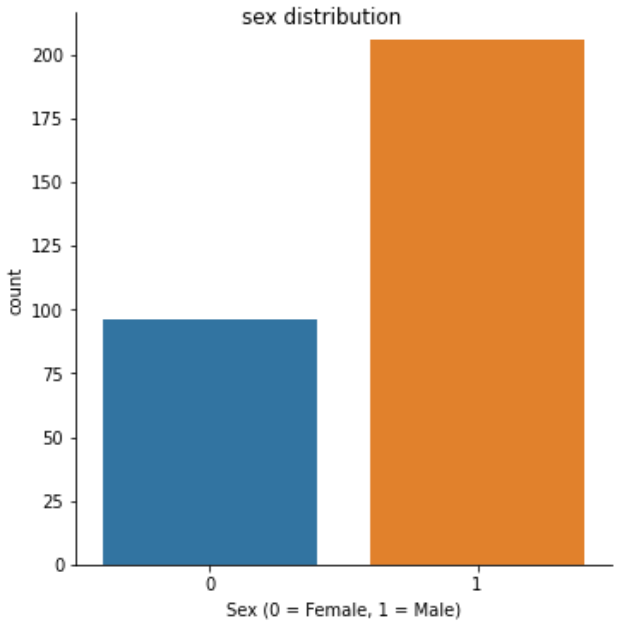
התפלגות המשתנים שבחרנו:

1) age – גיל המטופל



עבור משתנה זה ניתן לראות שגיל רוב המטופלים הוא 40-70, כנראה כי כשצעירים יותר מגיעים פחות לבית החולים, וכשמבוגרים יותר, יש פחות אנשים חיים בגילים אלו ולכן יש פחות ייצוג שלהם בבית החולים.

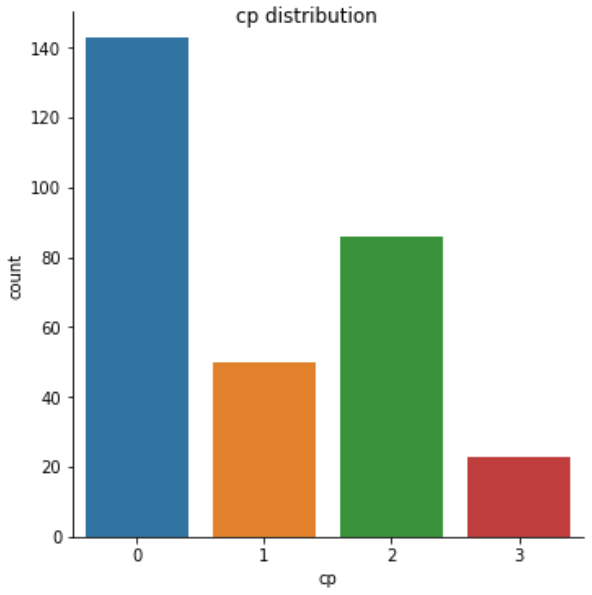
2) sex – מין המטופל



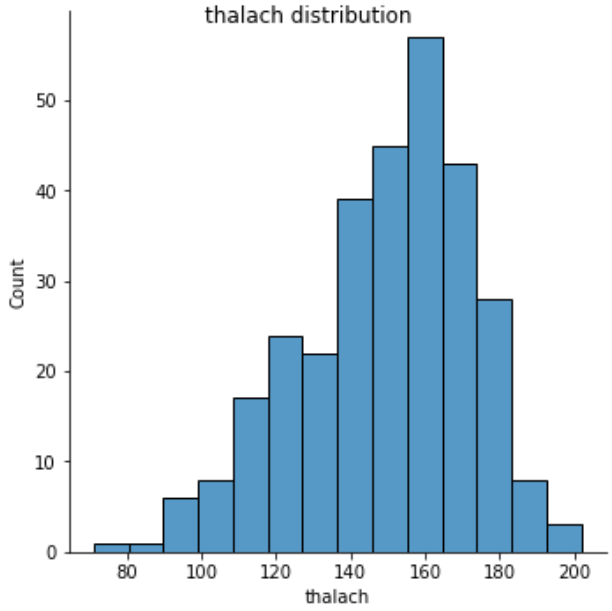
עבור משתנה זה ניתן לראות שיש פי 2 מטופלים גברים לעומת נשים.

3) cp – סוג כאב החזה

ניתן לראות שכאב החזה הנפוץ ביותר הוא מסוג "0" ואחריו סוג "2".

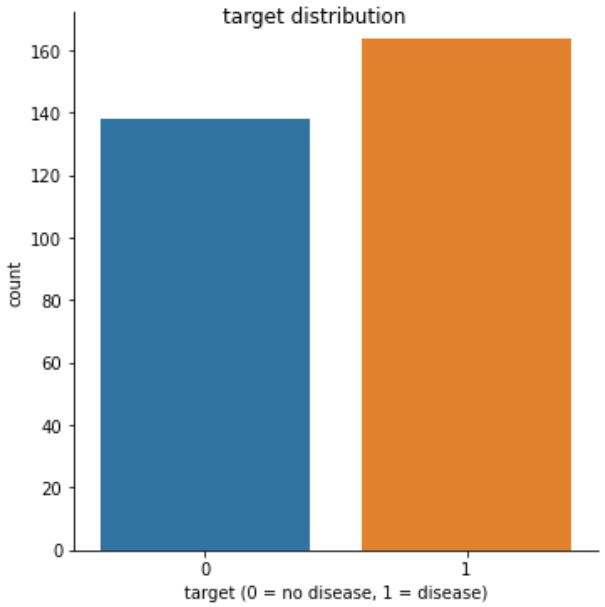


4) thalach – קצב דפיקות הלב המקסימלי



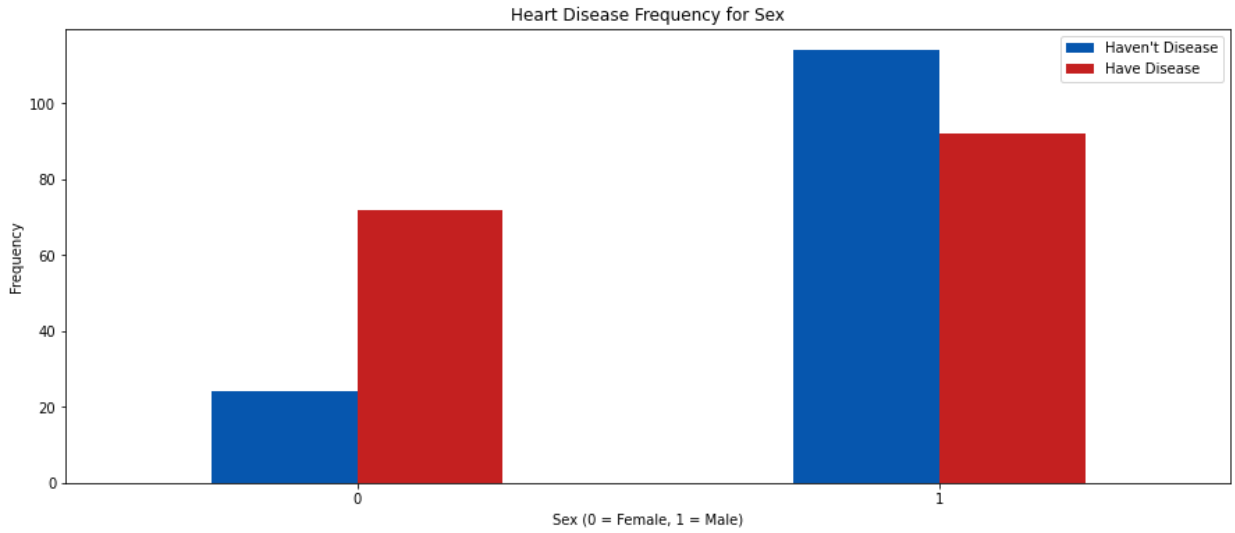
5) Target – האם חולה לב או לא

ניתן לראות שרוב המטופלים הם חולי לב.

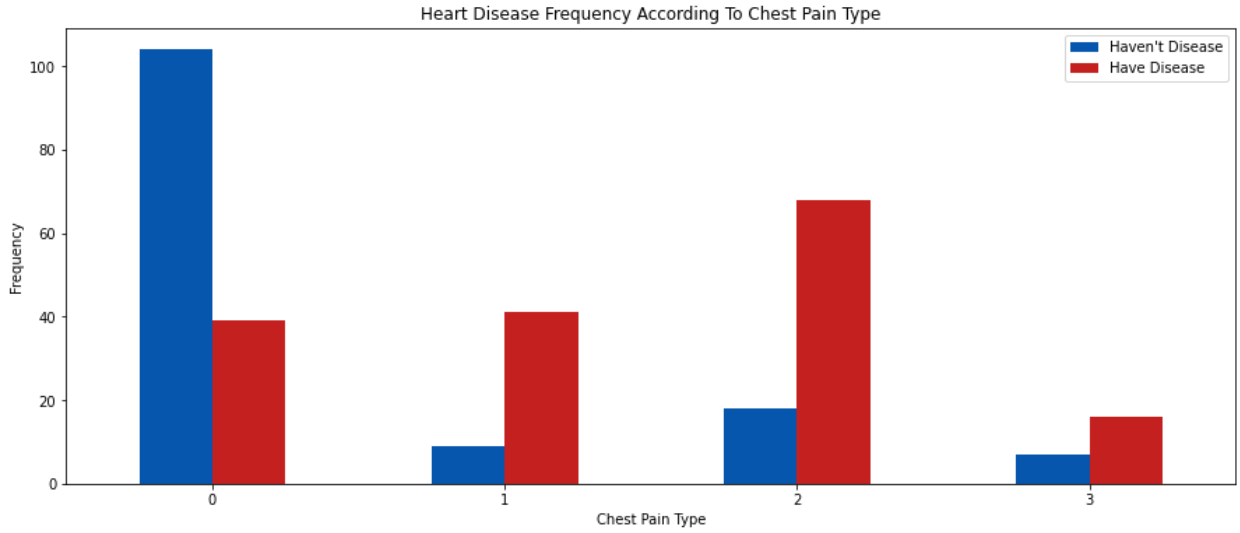


(שאלה 2 סעיף ב')

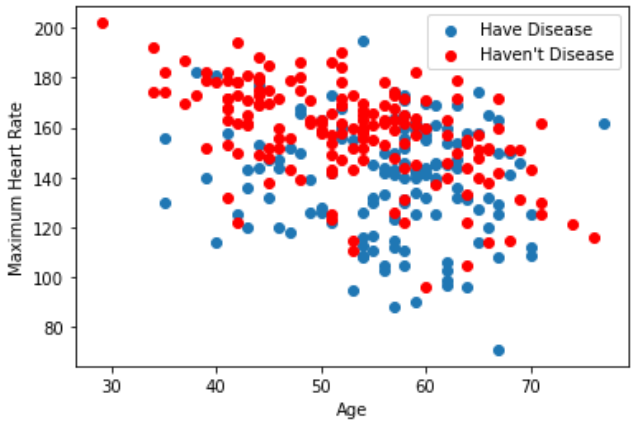
1) המשתנים בהם השתמשנו לגרף זה: sex, target.
בגרף הבא ניתן לראות את כמות האנשים בעלי מחלות הלב וכמות האנשים אשר אין להם מחלות לב, לפי המין של האדם. ניתן לראות כי אמנם ישנם יותר גברים במאגר הנתונים, אך מבחינת היחס בין אלו שיש להם מחלות לב לאלו שלא, בקרב שני המינים השונים, במאגר הנתונים רוב הנשים חולות לב ורוב הגברים הם אינם חולי לב.



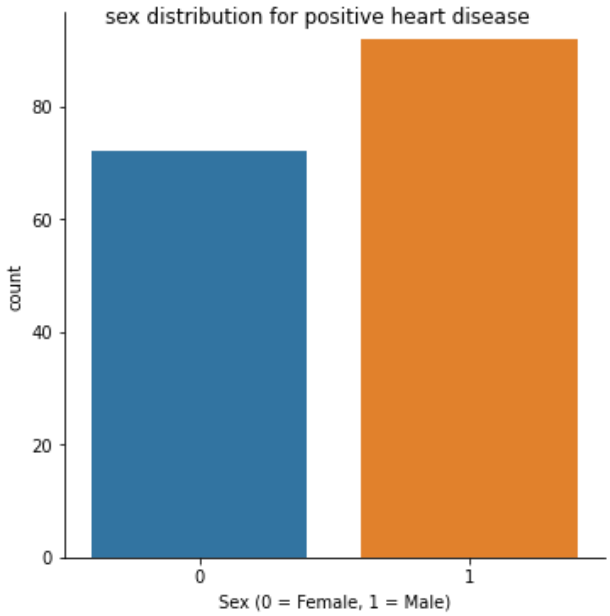
2) המשתנים בהם השתמשנו לגרף זה: cp, target.
בגרף הבא ניתן לראות את כמות סוגי כאב החזה אותם חוו המטופלים שבמאגר הנתונים, לפי האם קיימת מחלת לב או לא. ניתן לראות שבקרב המטופלים שאינם חולי לב, כאב חזה מסוג "0" הוא הנפוץ ביותר. עוד ניתן לראות כי עבור סוגי הכאב "1", "2", "3", רוב המטופלים שחוו כאבים מהסוגים הללו, הם בעלי מחלות לב.



3) המשתנים בהם השתמשנו לגרף זה: age, thalach, target.
בגרף הבא ניתן לראות את קצב דפיקות הלב המקסימלי של כל מטופל לפי גילו, כאשר יש הפרדה בין קבוצת המטופלים בעלי מחלות הלב לבין קבוצת המטופלים שאינם חולי לב. ניתן לראות שבקרב המטופלים חולי הלב, ישנו ריכוז גבוה יותר בחלקו העליון של ציר ה-y, כלומר - קצב הלב המקסימלי שלהם גבוה יותר.



Estimation and hypothesis testing (שאלה 3)



א) השאלה שנרצה לבדוק היא האם יש הבדל בגיל הממוצע של חולי הלב בין גברים לנשים. שאלה זו מעניינת אותנו מכיוון שעבור שנינו, על בסיס ידע קודם שהיה לנו, התשובה המיידית לשאלה זו הייתה שאכן יש הבדל, ורצינו לבחון האם הנתונים הללו מתיישבים עם מה שחשבנו.

<u>הערה:</u> בשלב זה נשתמש בדאטה חדש שמכיל אך ורק את המטופלים חולי הלב מתוך הדאטה המקורי.

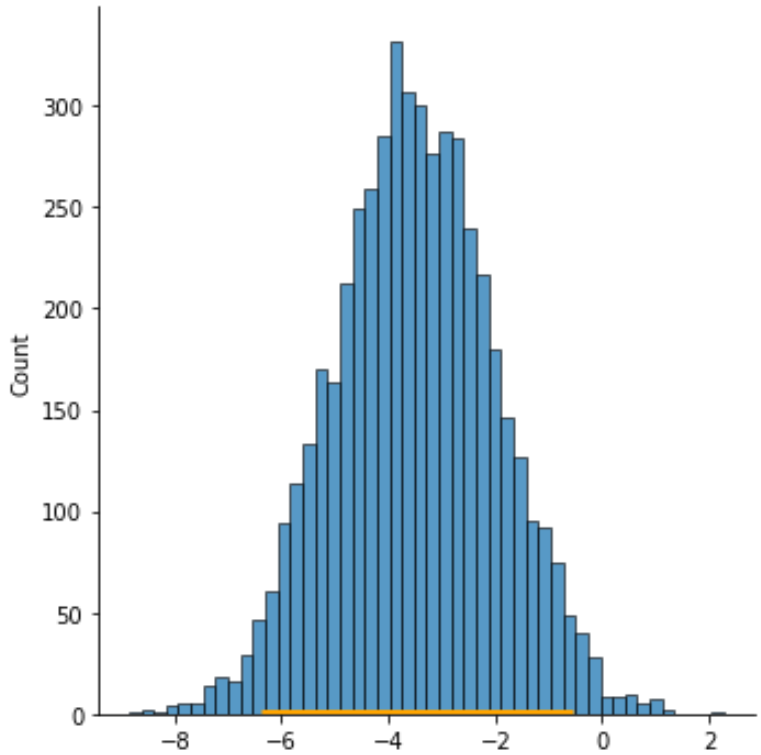
ב) H0: הגיל הממוצע של גברים חולי לב שווה לגיל הממוצע של נשים חולות לב.

H1: הגיל הממוצע של גברים חולי לב לא שווה לגיל הממוצע של נשים חולות לב.

<u>הערה:</u> בחרנו את השערת האפס להיות כך מכיוון שיהיה לנו הרבה יותר קל לסתור השערה ספציפית (אין הבדל) מאשר השערה כללית (קיים הבדל, לא יודעים בדיוק מה הוא...).

סטטיסטי המבחן שלנו הוא ההפרש בגיל הממוצע של חולי הלב בין גברים לנשים.

ג) <u>סוג האנליזה שעשינו:</u>
לדעתנו שיטת bootstrap היא המתאימה ביותר לבחון את מקרה זה מכיוון שלא נתונים לנו כל הנתונים על האוכלוסייה הכללית אלא רק מדגם (שהנחנו שהוא מייצג), והשיטה הנ"ל מאפשרת לייצר ממנו מדגמים רבים אשר מאפשרים לנו להסיק מסקנות (שבהסתברות גבוהה הן נכונות) על האוכלוסייה הכללית.



ראשית, נניח שמאגר הנתונים שקיבלנו מייצג בצורה טובה את האוכלוסייה הכללית. על מאגר הנתונים שקיבלנו הרצנו 5000 סימולציות אשר בכל אחת מהן יצרנו מדגם נוסף שנדגם באקראי מתוך מאגר הנתונים.
המדגם שיצרנו זהה בגודלו למאגר נתונים המקורי, ומכיוון שנדגם באקראי עם החזרה- יחס הגברים-נשים הינו בהסתברות גבוהה דומה ליחס במאגר הנתונים המקורי. בכל סימולציה חישבנו את ממוצע הגיל שבו גברים חולים במחלות לב וממוצע הגיל שבו נשים חולות במחלות לב, חישבנו את ההפרש בין הממוצעים והכנסנו להיסטוגרמה.

מסקנות:

ניתן לראות שרווח הסמך שיצא לנו הוא [-6.4, -0.57], כלומר שההפרש אפס לא נכנס ברווח הסמך ברמת מובהקות של 95% (המשמעות היא שב-95% מהפעמים שנחשב רווח סמך- אפס לא יהיה בתוכו),

לכן נוכל לדחות את השארת האפס האומרת שהגיל הממוצע של גברים חולי לב שווה לגיל הממוצע של נשים חולות לב (ברמת מובהקות של 95%).

מגבלות אפשריות:

1) במידה שמאגר הנתונים עליו עבדנו לא מייצג את האוכלוסייה הכללית, המסקנות אותן נסיק כאשר נשתמש ב-bootstrap יצאו לא מדויקות עבור העולם האמיתי, אלא רק עבור עולם היפותטי בו הנתונים שיש לנו מייצגים את האוכלוסייה הכללית בו.
2) המאגר עליו עבדנו מכיל בסך הכול כ-164 רשומות, לכן- מכיוון שאוכלוסיית חולי הלב יכולה להיות גדולה מאוד (אנו לא יודעים מה הוא הגודל האמיתי), יכול להיווצר מצב בו המדגם שנתון לנו קטן מדי (באופן יחסי), והמדגמים שנבחרים ממנו באקראי לא מייצגים בצורה הולמת את האוכלוסייה הכללית. לכן, המסקנות שנקבל על בסיסם עלולות להיות מוטות.

Prediction (שאלה 4)

נרצה לסווג האם מטופל הוא חולה לב או לא על סמך נתונים בעלי קורלציה גבוהה עם המשתנה target (האם יש מחלות לב או לא) מתוך הדאטה שלנו.

נרצה לאמן מסווג שכזה כדי לדעת לנהוג בהתאם כאשר מטופל הוא חולה לב, על מנת שיקבל את הטיפול ההולם עבור מצבו הבריאותי. מסווג זה עשוי להציל חיים כאשר לא ידוע האם למטופל יש מחלת לב, בכך שהוא נותן חיזוי ראשוני ודי מדויק על מצבו של המטופל.

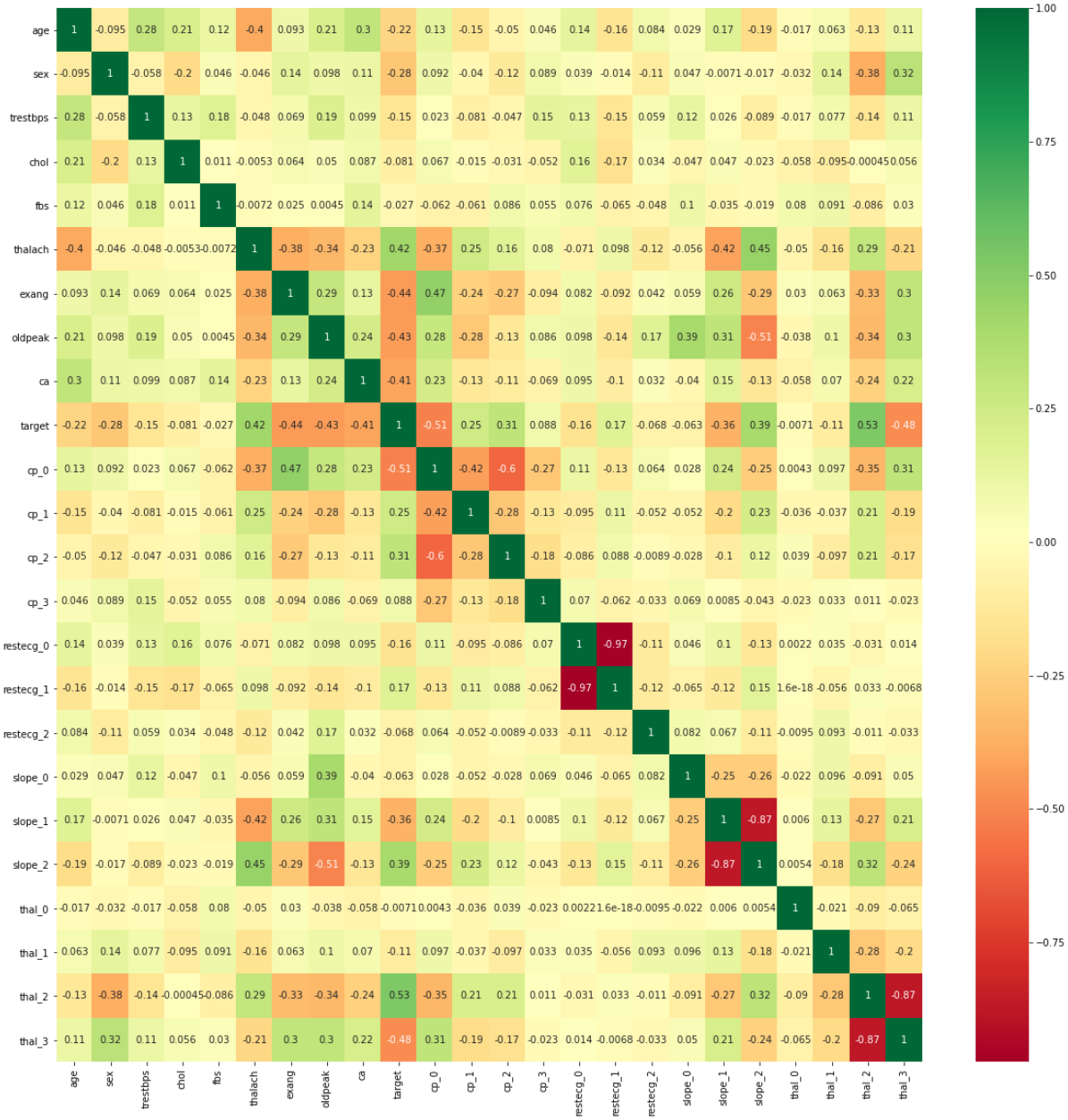
על מנת לעשות זאת, נצטרך לפצל משתנים קטגוריאליים למספר משתנים בינאריים. המשתנים אותם נפצל הם: cp, restecg, slope, thal.

נעשה השמה של הדאטה פריים לאחר הפיצול לתוך המשתנה encoded_df.

כעת, על מנת שלא יהיו לנו עיוותים הנוגעים לסדרי הגודל של המשתנים, נצטרך לעשות scaling ל-encoded_df.

נעשה השמה של הדאטה פריים לאחר ה-scaling לתוך המשתנה scaled_df.

כעת, לאחר הפיצול, נסתכל על טבלת החום הבאה שיצרנו:

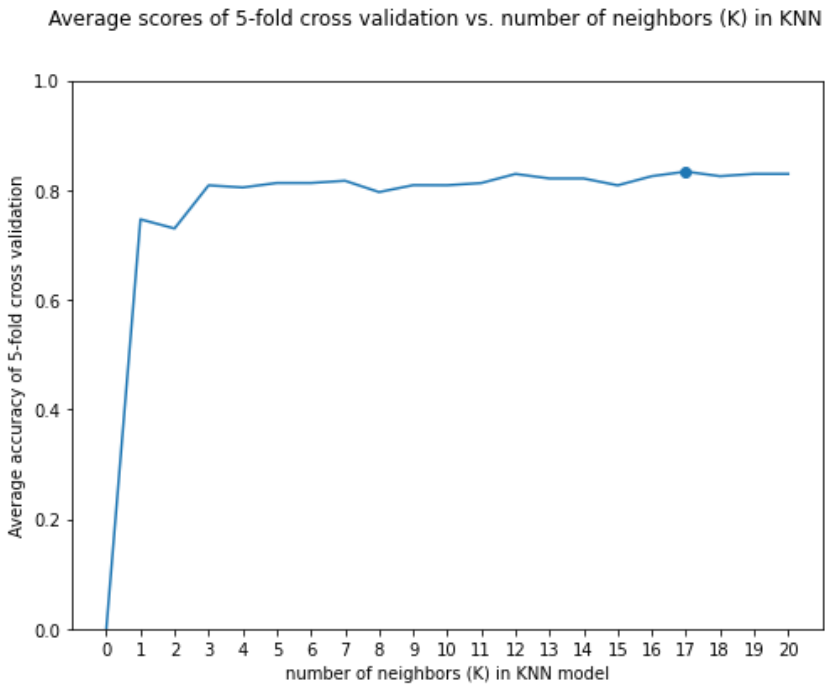


ניקח משתנים אשר בינם ובין המשתנה target ישנה קורלציה יחסית גבוהה. נקבע קורלציית מינימום להיות הערך 0.2 בערך מוחלט, (כלומר, 0.2 ומעלה או -0.2 ומטה), ואת אלו שלא עומדים בתנאי זה- נסיר.

המשתנים בהם בחרנו הם:

- Age
- Sex
- Thalach
- Exang
- Oldpeak
- Ca
- cp_0
- cp_1
- cp_2
- slope_1
- slope_2
- thal_2
- thal_3

ניקח את ערכי העמודות הנ"ל ונציב בתוך משתנה X. את ערכי עמודת target נציב בתוך משתנה Y.



נחלק את הדאטה שלנו באופן רנדומלי לקבוצת אימון וקבוצת מבחן.

בחרנו ש-20% מהרשומות יכנסו לקבוצת המבחן מכיוון שכך לדעתנו יש מספיק רשומות עליהן המסווג יתאמן ומספיק רשומות עליהן הוא ייבחן בסופו של דבר.

כעת, נשתמש ב-k-fold cross validation

ונבחר k=5, על האלגוריתם של KNN.

<u>הערה:</u> k-fold cross validation זו שיטה בה מחלקים את קבוצת האימון ל-k קבוצות שוות בגודלן. בכל שלב בוחרים קבוצה אחת מבין k הקבוצות שתשמש בקבוצת מבחן, והשאר יהיו קבוצות אימון. בכל שלב מאמנים את המודל ובוחנים על הקבוצה שנבחרה.

כך עושים על כל ערכי ה-K השונים, על מנת לבחור את הK (מספר השכנים) האופטימלי (זה שהדיוק שלו הוא המקסימלי).

נבחר לבדוק ערכי K שונים בין 1 ל-20 ולא מעבר על מנת להימנע מרעשי רקע העלולים להגרם כאשר מספר השכנים גדול יותר, מכיוון שאם נתייחס למספר רב יותר של שכנים, המסווג עלול לקחת בחשבון שכנים שהמרחק האוקלידי בין הנקודה הנבחנת, אליהם, הוא גדול באופן יחסי, ובכך להטות את התוצאות.

בהרצה הזו יצא לנו K=17, עם 83% דיוק.

כעת, לאחר שבחרנו את ה-K הנ"ל, נבדוק את אחוז הדיוק שלו על קבוצת המבחן.

קיבלנו שדיוק המסווג שלנו כאשר K=17 הוא 82%.

<u>מגבלות אפשרויות:</u>

1) מכיוון שהחלוקה לקבוצת אימון וקבוצת מבחן נעשית באופן רנדומלי, קיים סיכוי נמוך שהרשומות שנבחרות לקבוצת האימון יהיו שונות מהרשומות שנבחרו לקבוצת המבחן, ולכן, המסווג שאימנו על קבוצת האימון לא יסווג בהצלחה את הרשומות של קבוצת המבחן.
לדוגמה: לקבוצת האימון ייבחרו רק כאלה עם ערכי cp מסוימים, ולקבוצת המבחן יבחרו רק כאלה עם ערך cp ספציפי. במצב זה המסווג יתקשה יותר בזיהוי הרשומות מקבוצת המבחן.

2) בשלב ה-Scaling, שיטת הנרמול בה השתמשנו רגישה יחסית לערכים קיצוניים. דבר זה עלול לפגוע בדיוק של המסווג מכיוון שהפרופורציות עלולות להיות מעוותת מה שיבוא לידי ביטוי בחישוב המרחק האוקלידי שהמסווג עושה ובכך יתקבלו תוצאות מוטות.

כיוונים עתידיים

שאלות שעלו לנו תוך כדי שלא ניתן לענות עליהן בעזרת הנתונים הנוכחיים:

1) במידה שהייתה עמודה המציינת באיזו מדינה גדל אותו מטופל, היינו יכולים לבחון האם יש קשר בין המדינה בה אדם גדל לסיכוי שלו להיות חולה לב.

2) במידה שהיו לנו עמודות על אורח החיים של המטופלים (לדוגמה: האם המטופל מעשן/עושה ספורט וכו'), יכולנו לבחון האם יש קשר בין אורח החיים של המטופל לבין הסיכוי שלו להיות חולה לב.